A Syntactic Parser with Semantic Filtering for Biomedical Text

Gondy Leroy[1], Thomas C. Rindflesch[2], Bisharah Libbus[2],
Halil Kilicoglu[2], Hsinchun Chen[3]

Claremont Graduate University[1]
National Library of Medicine[2]
University of Arizona[3]

Because of the large volume of online literature in biomedicine, potentially useful information is often underutilized by researchers. Natural language processing techniques are increasingly used for enhanced access to this literature, often extracting specific information on genes, proteins, and other phenomena, including relationships such as activation and inhibition. Some extraction systems use pattern matching or underspecified syntactic methods to yield high recall, while others employ semantic processing aimed at a narrowly focused target to produce high precision.

We are experimenting with preliminary syntactic parsing of biomedical text followed by semantic filtering to combine the strongest features of both approaches. The Genescene parser [1] identifies syntactic predications that have simple noun phrases as arguments and verbs or prepositions as predicates. A wide range of relations are identified; however, semantic labeling of the predications is not provided. SemGen [2] extracts semantic relationships focused on the etiology of genetic diseases and uses the Unified Medical Language System® Metathesaurus® and other resources [3] to label noun phrases as either genetic phenomena or disorders.

In a pilot study, thirty-seven MEDLINE citations on the genetic disorder Crohn's disease were processed by the Genescene syntactic parser, and a total of 437 relationships were identified. These predications were then subjected to semantic filtering. For example, from the text *In human monocytes, infliximab enhanced TNF-alpha gene expression…*, syntactic processing identified the relation "infliximab - enhance - TNF-alpha gene expression," and SemGen labeled both phrases as genetic phenomena.

We evaluated SemGen's accuracy in processing the 781 noun phrases identified by the Genescene parser. Eleven percent were labeled as a disorder and 17% as a genetic phenomenon. Overall precision was 89% and recall was 84%. Seventy-four percent of the relationships relevant to the genetic etiology of Crohn's disease had at least one argument labeled correctly.

We believe that this methodology shows considerable promise. Syntactic processing can be used effectively to construct a comprehensive knowledge base, which can then be focused on a particular topic by subsequent semantic filtering. Several different filters could be applied, allowing a researcher to view only selected aspects of the general topic.

1. Leroy G, Chen H, Martinez JD. A shallow parser based on closed-class words to capture relations in biomedical text. J Biomed Inform. In press.
2. Rindflesch TC, Libbus B, Hristovski D, Aronson AR, Kilicoglu H. Semantic relations asserting the etiology of genetic diseases. Proc AMIA Symp. In press.
3. Tanabe L, Wilbur WJ. Tagging gene and protein names in biomedical text. Bioinformatics. 2002 Aug;18(8):1124-32.